HY562 – Προχωρημένα Θέματα Βάσεων Δεδομένων

Εαρινό εξάμηνο 2009-2010

Διδάσκοντες: Νίκος Σπυράτος, Γιάννης Τζίτζικας Βασίλης Χριστοφίδης

Ώρες: Δευτέρα 11-1 Β211, Τρίτη 3-5 Β211, Πέμπτη 11-1 ΡΑ201

(το πρώτο μάθημα θα γίνει Τρίτη,
κάποια από τα επόμενα θα γίνονται Δευτέρα αντί για Τρίτη)

Περιγραφή Μαθήματος

Σκοπός του Μαθήματος

Σκοπός του μαθήματος είναι η εισαγωγή των φοιτητών στις βασικές έννοιες και τεχνικές για την ενοποίηση/ολοκλήρωση δεδομένων και τη διαχείριση (εξόρυξη γνώσης, διαβάθμιση, εξερεύνηση) μεγάλων όγκων πληροφοριών. Οι φοιτητές θα έρθουν σε επαφή με σύγχρονα ερευνητικά θέματα και προβλήματα και θα δοθεί έμφαση στη συγκριτική αξιολόγηση και στην ανάπτυξη της κριτικής σκέψης.

Σκεπτικό

Αποτελεί βασική απαίτηση σήμερα η ενοποίηση/ολοκλήρωση της πληροφορίας. Η παροχή ενιαίας πρόσβασης σε πολλές πηγές απαιτεί τη γεφύρωση διαφόρων τύπων ετερογένειας και το πρόβλημα είναι δύσκολο αφού η φύση των δεδομένων αλλάζει και ο όγκος τους αυξάνεται αλματωδώς. Η συγκέντρωση, είτε με εικονικό (virtual) ή με υλοποιημένο (materialized) τρόπο μεγάλου όγκου πληροφοριών κάνει επιτακτική την ανάπτυξη προηγμένων τεχνικών για τη διαχείρισή τους. Για παράδειγμα, η εξόρυξη πληροφορίας από μεγάλους όγκους δεδομένων αποτελεί πλέον βασική λειτουργικότητα (data mining, inductive reasoning) με αποτέλεσμα τη δημιουργία δύο νέων ερευνητικών κατευθύνσεων στην περιοχή των βάσεων δεδομένων: (α) των αποθηκών δεδομένων (data warehouses) για τη συσσώρευση μεγάλων όγκων δεδομένων και την κατάλληλη οργάνωση τους για ανάλυση δεδομένων και εξόρυξη πληροφορίας, και (β) την επέκταση των γλωσσών επερώτησης ώστε να υποστηρίζουν διαβάθμιση ως προς διάφορα κριτήρια (στατιστικά, προτιμήσεις ή άλλα) με αποτελεσματικό και αποδοτικό τρόπο. Για το λόγο αυτό τα τελευταία χρόνια υπάρχει προσπάθεια εισαγωγής στο χώρο των βάσεων δεδομένων εννοιών και τεχνικών που παραδοσιακά εντάσσονταν στην περιοχή της μηχανικής μάθησης (machine learning), της στατιστικής ανάλυσης (statistical analysis) και της Ανάκτησης Πληροφορίας (Information Retrieval).

Περιεχόμενο

Αρχικά γίνεται μια εισαγωγή/επανάληψη σε θεμελιώδεις γνώσεις υποβάθρου οι οποίες είναι απαραίτητες για την κατανόηση πολλών θεμάτων και χρήσιμες για κάποιον που θέλει να κάνει έρευνα στον χώρο των βάσεων δεδομένων. Συγκεκριμένα γίνεται επανάληψη βασικών εννοιών των διακριτών μαθηματικών (Set Theory, Order Theory, Closures, Fixed Points, Lattices) και του σχεσιακού μοντέλου, και κατόπιν γίνεται εισαγωγή στα βασικά ζητήματα της μηχανικής επερωτήσεων (εγκλεισμός, ισοδυναμία, ελαχιστοποίηση επερωτήσεων και επέκταση επερωτήσεων με αναδρομή).

Εν συνεχεία το μάθημα θα επικεντρωθεί σε τρεις θεματικές περιοχές.

Ενότητα "Ενοποίηση Πληροφορίας και Ανάκτηση Πληροφοριών (Information Retrieval)

- Ανάκτηση Πληροφοριών και Ολοκλήρωση Πληροφορίας
- Επιλογή Πηγής (source selection) και Συνάθροιση Απαντήσεων
    - Συνάθροιση Διατάξεων (rank aggregation)
    - Γρήγορη Εύρεση των Κορυφαίων Στοιχείων μιας (Συναθροισμένης) Διάταξης (αλγόριθμοι top-K)
- Αυτόματη Διαβάθμιση Πλειάδων Απάντησης
- Γρήγορη Εύρεση Μέγιστων Στοιχείων (skylines)
- Διαβάθμιση Πλειάδων Απάντησης βάσει Προτιμήσεων

Ενότητα «Αποθήκες Δεδομένων» (Data Warehouses)

- Introduction: From transactional databases to data warehouses
- Functional databases
    - Functional Schema and functional algebra
    - Path expressions and OLAP queries
    - Query optimization
    - Mapping to relational
    - Report generation, SQL support (extensions)
    - Physical considerations (indexing)
- Knowledge discovery in data warehouses
    - Classification, clustering, association rules
- View management

Ενότητα «Μεσολαβητές» (Mediators)

- Virtual versus Materialized Integration
- The notion of Mapping
- LAV, GAV and combinations
- Mappings and Query Evaluation
- Case studies

Βαθμολογία

Βαθμός = 70% Βαθμός Τελικής Εξέταση + 30% Βαθμός Εργασίας.

Η Εργασία μπορεί να είναι (α) προγραμματιστική, ή (β) παρουσίαση άρθρων στην τάξη και σύνταξη σχετικής αναφοράς. Στη (β) περίπτωση η γραπτή αναφορά μετράει 10% του βαθμού ενώ η παρουσίαση στην τάξη 20%.